





BMJ Open COVID-19 surveillance data quality issues: a national consecutive case series

Cristina Costa-Santos ^{1,2} Ana Luisa Neves ^{1,2,3} Ricardo Correia,^{1,2} Paulo Santos ^{1,2} Matilde Monteiro-Soares ^{1,2,4} Alberto Freitas,^{1,2} Ines Ribeiro-Vaz,^{1,2,5} Teresa S Henriques,^{1,2} Pedro Pereira Rodrigues,^{1,2} Altamiro Costa-Pereira,^{1,2} Ana Margarida Pereira,^{1,2} Joao A Fonseca^{1,2}

To cite: Costa-Santos C, Neves AL, Correia R, *et al*. COVID-19 surveillance data quality issues: a national consecutive case series. *BMJ Open* 2021;**11**:e047623. doi:10.1136/bmjopen-2020-047623

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2020-047623>).

AMP and JAF contributed equally.

Received 10 December 2020
Accepted 05 November 2021



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Dr Cristina Costa-Santos;
csantos.cristina@gmail.com

ABSTRACT

Objectives High-quality data are crucial for guiding decision-making and practising evidence-based healthcare, especially if previous knowledge is lacking. Nevertheless, data quality frailties have been exposed worldwide during the current COVID-19 pandemic. Focusing on a major Portuguese epidemiological surveillance dataset, our study aims to assess COVID-19 data quality issues and suggest possible solutions.

Settings On 27 April 2020, the Portuguese Directorate-General of Health (DGS) made available a dataset (DGSApril) for researchers, upon request. On 4 August, an updated dataset (DGSAugust) was also obtained.

Participants All COVID-19-confirmed cases notified through the medical component of National System for Epidemiological Surveillance until end of June.

Primary and secondary outcome measures Data completeness and consistency.

Results DGSAugust has not followed the data format and variables as DGSApril and a significant number of missing data and inconsistencies were found (eg, 4075 cases from the DGSApril were apparently not included in DGSAugust). Several variables also showed a low degree of completeness and/or changed their values from one dataset to another (eg, the variable 'underlying conditions' had more than half of cases showing different information between datasets). There were also significant inconsistencies between the number of cases and deaths due to COVID-19 shown in DGSAugust and by the DGS reports publicly provided daily.

Conclusions Important quality issues of the Portuguese COVID-19 surveillance datasets were described. These issues can limit surveillance data usability to inform good decisions and perform useful research. Major improvements in surveillance datasets are therefore urgently needed—for example, simplification of data entry processes, constant monitoring of data, and increased training and awareness of healthcare providers—as low data quality may lead to a deficient pandemic control.

INTRODUCTION

The availability of accurate data in an epidemic is crucial to guide public health measures and policies.¹ During outbreaks, making epidemiological data openly available, in real time, allows researchers with different backgrounds to use diverse

Strengths and limitations of this study

- As accurate data in an epidemic are crucial to guide public health policies, this study identifies quality issues of the COVID-19 surveillance datasets.
- Only studied the quality issues of COVID-19 surveillance datasets from one country, Portugal.
- Several strategies for improving quality of health-care data were recommended.

analytical methods to build evidence^{2 3} in a fast and efficient way. This evidence can then be used to support adequate decision-making which is one of the goals of epidemiological surveillance systems.⁴ To ensure that high-quality data are collected and stored, several factors are needed, including robust information systems that promote reliable data collection,⁵ adequate and clear methods for data collection and integration from different sources, as well as strategic data curation procedures. Epidemiological surveillance systems need to be designed having data quality as a high priority and thus promoting, rather than relying on, users' efforts to ensure data quality.⁶ Only timely, high-quality data can provide valid and useful evidence for decision-making and pandemic management. On the contrary, using datasets without carefully examining the metadata and documentation that describes the overall context of data can be harmful.⁷

The low data quality of epidemiological surveillance systems has been a matter of concern worldwide. In fact, Boes and colleagues assessed the German surveillance system for acute hepatitis B infections. They concluded that although timeliness improved over the evaluation period, data quality in terms of completeness of information decreased considerably. Authors also stress that as improved data completeness is required to adequately design prevention activities, reasons for this decrease should

further be explored.⁸ On the other hand, other authors assessed timeliness and data quality of Italy's surveillance system for acute viral hepatitis and concluded that this system collects high-quality data, but wide reporting delays exist.⁹ Another study evaluated the quality of the influenza-like illness surveillance system in Tunisia and concluded that to better monitor influenza, the quality of data collected by this system should be closely monitored and improved.¹⁰ Visa and colleagues, in Nigeria, evaluated the Kano State malaria surveillance system and recommended strategies to improve data quality.¹¹ Regarding COVID-19 pandemic, a recent study that evaluated the accuracy of COVID-19 data collection by the Chinese Center for Disease Control and Prevention, WHO, and European Centre for Disease Prevention and Control showed noticeable and increasing measurement errors in the three datasets as more countries contributed data for the official repositories.⁷

At the moment, producing these high-quality datasets within a pandemic is nearly impossible without a broad collaboration between health authorities, health professionals and researchers from different fields. The urgency to produce scientific evidence to manage the COVID-19 pandemic contributes to lower quality datasets that may jeopardise the validity of results, generating biased evidence. The potential consequences are suboptimal decision-making or even not using data at all to drive decisions. Methodological challenges associated with analysing COVID-19 data during the pandemic, including access to high-quality health data, have been recognised¹² and some data quality concerns were described.⁷ Nevertheless, to our knowledge, there is no study performing a structured assessment of data quality issues from the datasets provided by the National Surveillance Systems for research purposes during the COVID-19 pandemic. Although this is a worldwide concern, this study will use Portuguese data as a case study.

The Portuguese systems to input COVID-19 data and the data flows

In early March, the first cases of COVID-19 were diagnosed in Portugal.¹³ The Portuguese surveillance system for mandatory reporting of communicable diseases is named SINAVE (National System for Epidemiological Surveillance) and is in the dependence of the Directorate-General of Health (DGS). COVID-19 is included in the list of mandatory communicable diseases to be notified through this system either by medical doctors (through SINAVE MED) or laboratories (SINAVE LAB). A COVID-19-specific platform (Trace COVID-19) was created for the clinical management of patients with COVID-19 and contact tracing. However, data from both SINAVE and Trace COVID-19 are not integrated in the electronic health record (EHR). Thus, healthcare professionals need to register similar data, several times, for the same suspect or confirmed case of COVID-19, increasing the burden of healthcare professionals and potentially leading to data entry errors and missing data. The SINAVE notification

form includes a high number of variables, with few or no features to help data input. Some examples include: (1) within general demographic characteristics, patient occupation is chosen from a drop-down list with hundreds of options and with no free text available; (2) the 15 questions regarding individual symptoms need to be individually filled using a three-response option drop-down list, even for asymptomatic patients; (3) in the presence of at least one comorbidity, 10 specific questions on comorbidities need to be filled; and (4) there are over 20 questions to characterise clinical findings, disease severity, and use of healthcare resources, including details on hospital isolation. Other examples of the suboptimal design are (5) the inclusion of two questions on autopsy findings among symptoms and clinical signs, although no previous question ascertains if the patient has died; (6) lack of a specific question on disease outcome (only hospital discharge date); (7) lack of validation rules that allow, for example, to have a disease diagnosis prior to birth date or to be discharged before the date of hospital admission; and (8) no mandatory data fields, allowing the user to proceed without completing any data. Furthermore, a global assessment of disease severity is included with the options 'unknown', 'severe', 'moderate' and 'not applicable' without a readily available definition and without the possibility to classify the disease as mild. This unfriendly system may impair the quality of COVID-19 surveillance data. The problems described have existed for a long time at SINAVE and they are usually solved by personal contact with the health local authorities. However, in the current COVID-19 pandemic scenario, and due to the pressure of the huge number of new cases reported daily, this does not happen at this moment.

There is more than one possible data flow from the moment the data are introduced until the dataset is made available to researchers. Figure 1 is an example of the information flow from data introduced by public health professionals until the analysis of data.

Since the beginning of the pandemic, several research groups in Portugal stated their willingness to contribute by producing knowledge and improving data systems and data quality.¹⁴ Researchers requested access to healthcare-disaggregated data related to COVID-19, in order to timely produce scientific knowledge to help evidence-based decision-making during the pandemic. In April, DGS made a document publicly available with the description and metadata of the dataset to be provided and a form to be filled by researchers to request this dataset.^{15 16} A research protocol and a documented approval by an ethical committee were also necessary. A metadata document was available and the researchers knew what variables they would receive if they formally requested the dataset. The variables available did not include, for instance, clinical presentation or specific disease symptoms. The variable formats were described in the provided metadata but were not discussed or adjusted based on researchers' opinions or needs. In the metadata, the coded value list was described (eg, Y=Yes, N=No, Unk=Unknown) but

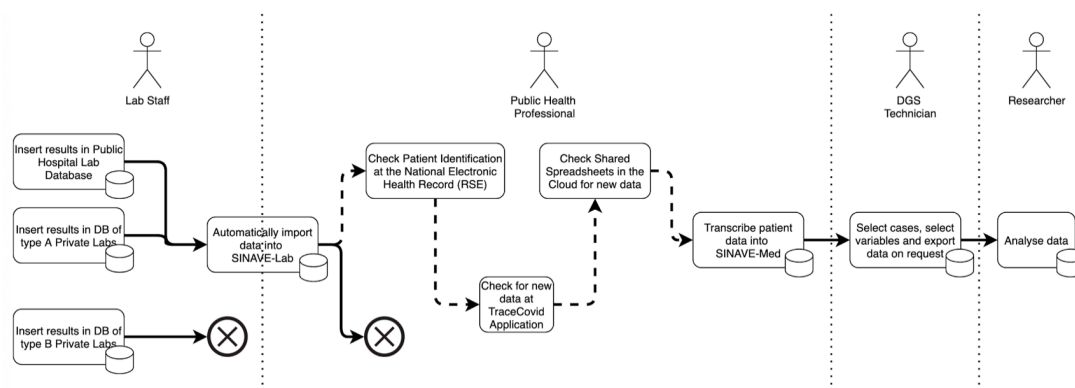


Figure 1 Example of one possible information flow from the moment the data are introduced until the dataset is made available to researchers. The ⊗ symbol means that data are not sent and therefore not present in the research database (DB). The dashed line represents a manual cumbersome process that is many times executed by public health professionals and that is very susceptible to errors. DGS, Directorate-General of Health.

not the coding mechanism, that is, how the form answer (given by the healthcare professional) was coded in the dataset. Therefore, although there was an ‘agreed’ dataset specification document, it was not complete enough to fully understand the provided data. Along with the data request form and metadata document, information was made available that researchers would receive weekly data updates.

On 27 April 2020, the DGS sent the described dataset (DGSApril) collected by the SINAVE MED and according to the metadata document made available before. At least 50 research groups received the data and started their dataset analyses. Weekly dataset updates were not provided and only on 4 August 2020, DGS sent an updated dataset (DGSAugust) to the research groups who had requested the first dataset, including COVID-19 cases already included in the initial dataset plus new cases diagnosed during May and June 2020. This updated dataset did not respect the metadata document initially provided, and had an inconsistent manifest, including some variables presented in a different format or absent. For example, instead of a variable with the outcome of the patient, the second dataset presented two dates: death and recovery date; and this new version did not distinguish between dead due to COVID-19 and dead due to reasons. The updated dataset also used definitions (for example, variable age was defined as the age at the time of COVID-19 onset or as age at the time of COVID-19 notification, in the first and second datasets, respectively). Also the variable of preconditions had different categories. For example, the first dataset the variable comorbidities had the category ‘cardiac disease’ and in the updated version of the dataset, this category was not present. All these aspects raised concerns regarding the updated dataset used for replication of the analysis made using the first version of data and consequently some concerns regarding its use for valid research.

We aimed to assess data quality issues of COVID-19 surveillance data and suggest solutions to overcome them, using the Portuguese surveillance datasets as an example.

METHODS

The data provided by DGS included all COVID-19-confirmed cases notified through the SINAVE MED and, thus, excluding those only reported by laboratories (SINAVE LAB).

The DGSApril dataset was provided on 27 April 2020 and the updated one (DGSAugust) on 4 August 2020. The available variables in both datasets are described in online supplemental file 1.

There was a variable named ‘outcome’, with the information on the outcome of the case, present in DGSApril dataset that was not available in the DGSAugust dataset. On the other hand, there were also some variables (dead, recovery, diagnosis and discharge dates) present in DGSAugust dataset that were not available in the DGSApril dataset.

The quality of the data was assessed through the analysis of data completeness and consistency between the DGSApril and DGSAugust datasets. For data completeness evaluation, missing information was classified as ‘system missing’ when there was no information provided (blank cells) and as ‘coded as unknown’ when the information ‘unknown’ was coded. Considering the consistency, both datasets were compared in order to evaluate if the data quality increased with the update sent 4 months later. As many data entry errors could be avoided using an optimised information system, the potential data entry errors in DGSAugust were also described.

The main outcome measures were: the frequency of cases with missing information, the frequency of cases with unmatched information between the datasets and its update, and the frequency of cases with wrong data entry (considered impossible values) for each variable.

The number of COVID-19 cases and the number of deaths due to COVID-19 were also compared with the public daily report by Portuguese DGS.¹⁷ We highlight that it is not expected that the daily numbers of cases and deaths reported publicly were coincident with the numbers obtained in the datasets made available to researchers as these datasets included only the COVID-19

cases notified through the SINAVE MED (excluding those only reported by laboratories). However, the calculation of this difference is important to estimate the potential bias that data of these (DGSApril and DGSAugust) datasets, provided by DGS to researchers, may have. This comparison is only possible in the DGSAugust dataset as in the DGSApril dataset, the variable date of diagnosis was not available.

Statistical methods

Descriptive statistics are presented as absolute and relative frequencies.

Data handling and analyses were performed using IBM SPSS Statistics V.26 and R V.4.0.3.

Patient and public involvement

As this study used secondary data, it was not possible to involve the participants in the study, in the design or in the recruitment and conduct of the study. However, the results have been and will continue to be disseminated not only with DGS but with patients and the whole community through the media.

RESULTS

Cases included and omitted

From the 20 293 COVID-19 cases included in the DGSApril dataset, only 80% ($n=16\,218$) had the same unique case identifier in the DGSAugust dataset. There were 4075 cases in the DGSApril dataset that were not included in the DGSAugust dataset or, alternatively, had changed the unique case identifier. The DGSAugust dataset provided a total of 38 545 COVID-19 cases, including 22 327 that were not available in DGSApril dataset: 5713 diagnosed until 27 April but that presumably were not included in

the DGSApril dataset, 16 609 diagnosed after the period included in the DGSApril dataset and 5 cases with missing information on diagnosis date (figure 2).

Considering the 5713 cases made available only in the DGSAugust and diagnosed before 27 April that, presumably, were not included in the DGSApril dataset, the majority (58%) were diagnosed in the 2 weeks immediately prior to 27 April (the date on which this database was made available). However, 42% were diagnosed more than 2 weeks before the DGSApril dataset was made available (figure 2).

Data completeness of both datasets

Several variables showed a low degree of completeness. For example, two variables ('date of first positive laboratory result' and 'case required care in an intensive care unit') had more than 90% of cases with missing information in DGSApril dataset—coded as unknown or system missing. In the DGSAugust dataset, the variable 'case required care in an intensive care unit' reduced the proportion of incomplete information to 26% of system missing and no cases were coded as unknown. However, the variable 'date of first positive laboratory result' still had 90% system missing in the DGSAugust dataset. Table 1 provides detailed information about missing information for each available variable.

Data consistency between DGSApril and DGSAugust datasets

The consistency of the information for cases identified with the same unique case identifier in both datasets ($n=16\,218$) was further evaluated (figure 1).

Table 2 presents the number and percentage of cases with different information, for each variable.

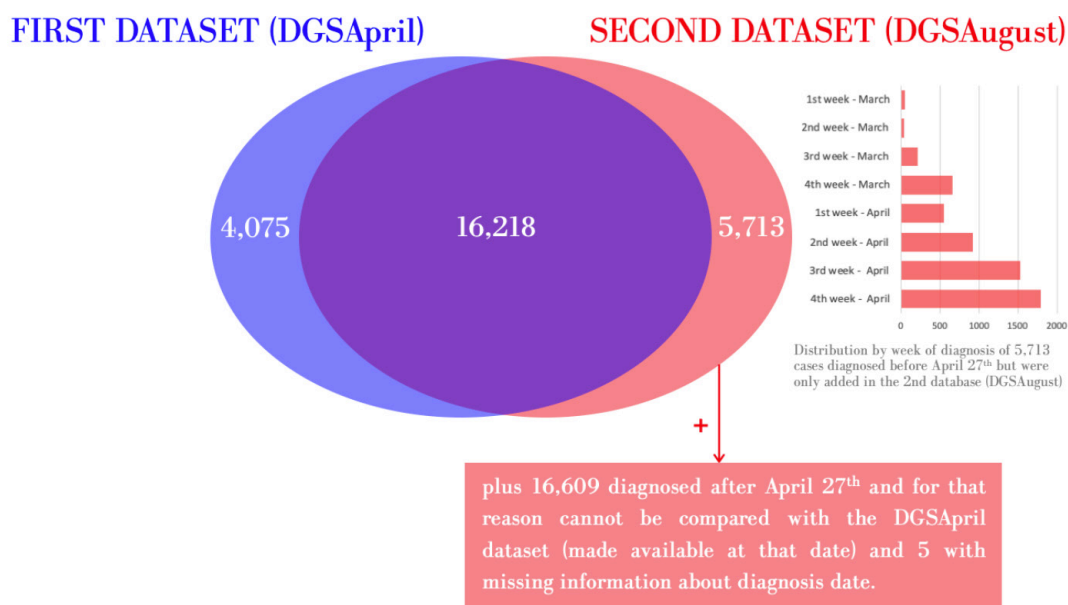


Figure 2 Number of unique case identifiers presented in the datasets of COVID-19 cases diagnosed since the start of the pandemic until 27 April (date when the first database was made available) and after 27 April. DGS, Directorate-General of Health.

Table 1 Data completeness (number and percentage of missing information) of each variable available in the DGSApril and DGSAugust datasets with COVID-19 cases provided by DGS

	DGSApril (n=20 293)		DGSAugust (n=38 545)	
	System missing n (%)	Coded as unknown n (%)	System missing n (%)	Coded as unknown n (%)
Unique case identifier (RecordID)	0	0	0	0
RecordID of the linked cases	*	*	*	*
Age	0	0	0	0
Probable place of infection	0	0	0	0
Gender	0	0	0	0
Hospitalisation	0	1623 (8)	3 (0)	3425 (9)
Outcome	0	23 (0)	†	†
Patient has underlying condition	0	2 (0)	15 407 (40)	2495 (6)
Date of first positive laboratory result	19 268 (95)	0	34 667 (90)	0
Date of diagnosis	‡	‡	7 (0)	0
Date of disease onset	4815 (24)	0	15 045 (39)	0
Date of death	‡	‡	37 390 (97)	0
Date of recovery	‡	‡	21 499 (56)	0
Only hospitalised cases	n=2973		n=4327	
Date of hospitalisation	386 (13)	0	860 (20)	0
Case required care in an intensive care unit	0	2712 (91)	1122 (26)	0
Level of respiratory support given to patient	0	1573 (53)	1364 (31)	172 (4)
Date of hospital discharge	‡	‡	3975 (92)	0

*Variable neither available in DGSApril dataset nor in DGSAugust dataset but described in the metadata file provided by DGS.

†Variable available in DGSApril dataset and described in the metadata file provided by DGS but not provided in DGSAugust dataset.

‡Variable neither available in DGSApril dataset nor described in the metadata file provided by DGS but provided in DGSAugust dataset. DGS, Directorate-General of Health.

Since the beginning of the pandemic, the report of COVID-19 within SINAVE kept the same data structure. A few variables related to specific symptoms that were progressively described as being in relation to COVID-19 infection were added (eg, anosmia or dysgeusia), but these variables (symptoms) were not included in the

analysed datasets (DGSApril and DGSAugust). However, some inconsistencies may be due to differences in the data format made available to researchers. Anyway, due to the lack of metadata information related to DGSAugust, it is not possible to harmonise such inconsistencies in data analysis. Some inconsistencies may be due to the update of the data made meanwhile, however many inconsistencies are difficult to understand because there is often information filled in the first dataset that is not filled in the updated dataset.

The variable 'underlying conditions' was the one showing a higher percentage of inconsistencies between both datasets, with more than half of cases showing different information when comparing the information from both datasets (table 2). Most of the inconsistencies were due to the cases recorded as 'no underlying conditions' in the DGSApril dataset and corrected to 'unknown' if the case has underlying conditions' or 'missing' in the updated dataset (DGSAugust) (42%, n=6851). There were 1952 cases (12%) recorded as 'no underlying conditions' in the first dataset and corrected to 'yes—underlying conditions' in the second one. There were also 99 (1%) cases with underlying conditions in the first dataset corrected to 'no underlying conditions' in the second one.

Table 2 Number and percentage of COVID-19 cases presented in both datasets (n=16 218) with information that did not match for each variable

	Healthcare data inconsistencies n (%)
Patient has underlying condition	8902 (55)
Age*	8326 (51)
Hospitalisation	253 (16)
Date of disease onset	2008 (12)
Date of first positive laboratory result	962 (6)
Probable place of infection	46 (0)
Gender of the reported case	1 (0)

*The definition of 'age' was different in both datasets: in DGSApril is the age at the time of COVID-19 onset, and in DGSAugust, the age at the time of COVID-19 notification.

The variable 'age' also had more than half of cases showing different information when comparing the information from both datasets (table 2). The difference in all cases with different information, except one, was 1 year old. The definition of 'age' was different in both datasets: in DGSApril is the age at the time of COVID-19 onset and, in DGSAugust, the age at the time of COVID-19 notification.

The variable 'hospitalisation' had 16% of cases (n=253) with unmatched information (table 2). One hundred and twenty-five cases were recorded as 'unknown if the case was hospitalised' in the DGSApril dataset and corrected to 'no hospitalisation' in the DGSAugust. Sixty-two cases were recorded as 'no hospitalisation' and corrected to 'hospitalised' or 'unknown information' in DGSApril and DGSAugust datasets, respectively. Fifty-five cases were recorded as hospitalised patients and corrected to 'no hospitalisation' or 'unknown information' in DGSApril and DGSAugust datasets, respectively. Only 11 cases changed from 'unknown if the case was hospitalised' to 'hospitalisation'.

The variable 'date of disease onset' had 12% of cases (n=2008) with unmatched information (table 2). In 1445 cases, information about the date of disease onset was provided only in DGSApril and 563 cases had dates in both datasets but the dates did not match.

The variable 'date of the first positive laboratory result' did not match in both datasets in 6% of the cases (n=962). In 5 cases, there was a date available in both datasets but the dates did not match; in 74 cases, the date was available only in the DGSApril dataset; and in 883 cases, the date was available only in the DGSAugust dataset.

The variable patient outcome (variable 'outcome') was not present in the DGSAugust dataset which instead presents the variables 'date of recovery' and 'date of death' (not presented in DGSApril) (table 1). In the DGSApril dataset, there were 1134 cases coded as 'alive, recovered and cured', but only 83% of those (n=947) had recovery date in the updated dataset (DGSAugust), which may be due to the lack of information on a specific date, despite knowing that the case result is alive, recovered and cured. In fact, 177 patients recorded as 'alive, recovered and cured' in the DGSApril did not have any date in the DGSAugust dataset. However, 10 patients recorded as 'alive, recovered and cured' in the DGSApril had a date

of death in the DGSAugust dataset. Seven of these were dates of death before April 19, which is incongruent. Among the 455 cases coded as 'died because of COVID-19' in the DGSApril dataset, 7 (2%) did not have a date of death in the second dataset.

Data entry errors in the updated dataset (DGSAugust)

The age of one patient is probably wrong (more than 130 years old). There were also male patients and elderly women registered as pregnant. There was a wrong diagnosis date (50-05-2020) and 19 patients had registered dates of diagnosis before the first official case of COVID-19 was diagnosed in Portugal. There were also two patients with a negative length of stay in hospital.

Of the 38545 cases included in the dataset, 6772 had recorded in the recovery date variable 'April 3', 1032 cases had recorded in the recovery date variable 'May 25' and 242 cases 'May 26'. The remaining 30499 cases had no information registered in this variable.

Number of COVID-19 cases and deaths provided by DGSAugust dataset and by daily public report

Table 3 shows the number of COVID-19 cases and deaths due to COVID-19 reported by DGSAugust dataset and by the daily public report. The DGSAugust dataset included 38 520 COVID-19 cases diagnosed between March and June, less 4003 cases (9%) than the daily public report provided by Portuguese DGS. However, when looking at data from March, the DGSAugust dataset reported more 669 cases (8%) than the daily public report. In April, May and June, the DGS dataset reported less 17%, 8% and 12% of cases than the public report provided, respectively.

The DGSAugust dataset reported 1155 deaths due to COVID-19 until the end of June, less 424 cases (27%) than the daily public report provided by the Portuguese DGS. However, in March, the DGSAugust dataset reported more five deaths due to COVID-19 (3%) than the daily public report. In April, May and June, the DGS dataset reported less 8%, 49% and 100% of cases than the public report provided, respectively.

Bias estimation

The most important problem in the first dataset is the potential underestimation of comorbidities due to the misclassification of cases with the information unknown

Table 3 Number of COVID-19 cases and deaths due to COVID-19 reported by DGSAugust dataset and by the daily public report

Month	COVID-19 cases reported by:			Deaths due to COVID-19 reported by:		
	DGSAugust	Daily public report	Difference	DGSAugust	Daily public report	Difference
March	8920	8251	+669	192	187	+5
April	13838	16736	-2898	750	820	-70
May	7113	7713	-600	213	417	-204
June	8649	9823	-1174	0	155	-155

DGS, Directorate-General of Health.

Table 4 Prevalence estimation for each precondition by DGSApril (used in Nogueira and colleagues¹⁹ study) and by the updated dataset

Precondition	Nogueira and colleagues' study (DGSApril) Prevalence (95% CI)	Updated dataset (DGSAugust) Prevalence (95% CI)
Asthma	1.36 (1.20 to 1.53)	4.74 (4.44 to 5.08)
Cancer	3.01 (2.78 to 3.26)	5.45 (5.12 to 5.81)
Cardiac disease	0.27 (0.20 to 0.35)	–
Haematological disorder	1.08 (0.09 to 1.24)	2.00 (1.79 to 2.22)
Diabetes	5.64 (5.33 to 5.97)	12.3 (11.8 to 12.8)
HIV/other immune deficiencies	0.53 (0.43 to 0.64)	1.35 (1.18 to 1.54)
Kidney disorder	1.98 (1.79 to 2.18)	4.33 (4.02 to 4.65)
Liver disorder	0.53 (0.43 to 0.64)	1.27 (1.11 to 1.46)
Lung disorder	3.39 (3.15 to 3.65)	4.50 (4.19 to 4.82)
Neuromuscular disorder	3.92 (3.66 to 4.19)	3.50 (3.23 to 3.79)
At least one precondition	16.6 (16.1 to 17.1)	40.3 (39.7 to 41.0)

DGS, Directorate-General of Health.

about preconditions as 'absence of precondition'. To estimate the potential systematic error identified by Costa-Santos and colleagues¹⁸ presented in the study by Nogueira and colleagues¹⁹ who analysed the first dataset, we estimate the prevalence of each precondition with the first dataset (those presented in Nogueira and colleagues' study¹⁹) and with the second dataset (where the cases with unknown information about preconditions were classified as missing information for that variable and not as 'precondition absent').

As table 4 evidence, the first dataset (DGSApril) presented a bias in the prevalence estimation of almost all preconditions probably due to the misclassification of cases with the information unknown about preconditions as 'absence of precondition'. Almost all the comorbidities in the DGSApril were greatly underestimated relatively to the second dataset. Even in the updated dataset (DGSAugust), the prevalence of preconditions may be underestimated. Indeed, for example, the estimate of

the prevalence of asthma in the Portuguese population is 6.8% (95% CI 6.0% to 7.7%).²⁰ According to Quinaz Romana and colleagues, the percentage of people in the Portuguese population who have at least one precondition is 58%.²¹ The Portuguese population of people infected with COVID-19 is unlikely to have a lower prevalence of comorbidities than the Portuguese general population.

DISCUSSION

The production of scientific evidence to help manage the COVID-19 pandemic is an urgency worldwide. However, if the quality of datasets is low, the evidence produced may be inaccurate and, therefore, have limited applicability. This problem may be particularly critical when low-quality datasets provided by official organisations lead to the replication of biased conclusions in different studies.

The problem of using datasets with suboptimal quality for research purposes during the COVID-19 pandemic probably occurs in a large number of countries. This study, using the Portuguese surveillance data, reports a high number of inconsistencies and incompleteness of data that may interfere with scientific conclusions. To date, we could identify three scientific papers reporting analysis of these data^{19 22 23} that may have been affected by the low quality of the datasets.²¹ Table 5 presents data quality issues identified in the provided datasets and possible solutions.

The issue of 'missing' versus 'absent' variable coding seems to be present in the findings of Nogueira and colleagues' study.¹⁹ The reduction of the risk of death in relation with comorbidities observed in the analysis of the first dataset is underestimated if we assume that the updated dataset is the correct one.²¹ In fact, these cases were registered as having no underlying conditions in the first dataset but corrected in the second dataset to 'unknown if the case has underlying conditions' or system missing. This problem might be due to the way these data were collected and/or were recorded in the database sent to the researchers. In the form used to collect COVID-19 surveillance data, comorbidities are recorded one by one after a general question assessing the presence of any comorbidity and the field is not mandatory. From a clinical point of view, it might be enough to register only positive data perceived as relevant (eg, the presence of

Table 5 Most frequent data quality issues and possible solutions

Issues	Solutions
'Missing' versus 'absent' variable coding	Automatically code blank cells as system missing Simplification of data entry processes, reusing the data already in the system Data interoperability
Differences in cases included	Guarantee same unique case identifier by recording it in the registry database
Data (in)completeness	Determine a core of mandatory variables
Data (in)consistency	Maintain same variables (and respective definitions) along time
Data entry errors	Improve information system (by determining possible values and limits) Data monitored and tracked

a specific diagnosis, but not its absence), especially in a high-burden context as the ongoing pandemic. In the context of clinical research, however, the lack of registered comorbidity data cannot be interpreted as the absence of comorbidities. A similar bias can be found in the other two studies reporting analysis of DGSApril dataset.^{22 23}

Another data quality issue is related to the discrepancies in cases included in both datasets. In fact, only 80% of cases included in the DGSApril dataset had the same unique case identifier in the DGSAugust dataset and only 74% of cases diagnosed until 27 April included in DGSAugust had the same unique case identifier in the DGSApril. Alternatively, the unique case identifier had been changed. We do not know if the unique identifier is generated in each data download or if it is recorded in the database. This last option will be the safest. Moreover, until 19 June, it was not mandatory to fill in the national health service user number in order to have a standard unique patient identifier. That may have led to not identifying duplicate SINAVE MED entries for the same patient and increased the difficulty in adequately merging data from SINAVE LAB, SINAVE MED and other data sources.

The high percentage of incomplete data in several variables may also produce biases whose dimensions and directions are not possible to estimate. In fact, as our results showed, half of the variables available in the DGSAugust dataset had more than one-third of missing information. Furthermore, that dataset was already incomplete since it only provides COVID-19 cases from the medical component of SINAVE totalling 90% of the cases reported by health authorities until the end of June 2020.¹⁷ It is unclear, however, why the updated version of the dataset in March reported more 669 COVID-19 cases and more 5 deaths than the public report (which would be expected to be more complete). Moreover, there were no reported dates of deaths in June in DGSAugust dataset, despite the 155 deaths reported in the public report during this month.

The consistency of variables in different updates of datasets is also an important quality issue. In fact, our results show that the variable 'age' was calculated differently in the two datasets: in the DGSApril dataset it was the age at the time of COVID-19 onset and in the DGSAugust dataset it was age at the time of COVID-19 notification. Despite this change in definition, the difference of 1 year in half of the cases does not seem to be completely justified only by this fact, since the two dates should be relatively close. Still related to this problem of inconsistent information and variables, we realised that some information may have been lost in the second dataset sent (DGSAugust). In fact, the outcome of the COVID-19 case is not presented in the second dataset. DGSAugust dataset only presents the recovery and death dates. It would be possible to reconstruct in the second dataset some of the information on the outcome variable presented in the first one. However, it would only be possible to directly recode those with 'date of recovery' as 'alive, recovered

and cured'; all other categories ('died of COVID-19'; 'died of other cause'; 'cause of death unknown'; 'still on medical treatment') are impossible to obtain from the dates of recovery or death. In fact, using only the variable 'date of death', it is not possible to determine if the patient died because of COVID-19, died of another cause or if the cause of death is unknown as in the DGSApril dataset. Moreover, 17% of cases coded as 'alive, recovered and cured' in the first dataset did not have the variable 'date of recovery' filled in the updated one. While the recovery date (when available) can be used as a proxy of the patient outcome, if this date is unknown in spite of a known recovery, we miss the whole outcome information.

In fact, in the DGSAugust dataset, it is assumed that the missing information about the recovery date implies that the case had not recovered yet. Also, the 'recovery date' had only three dates even though it refers to a 4-month period.

All the described errors, inconsistencies, data incompleteness, changes in the variables' definitions and format may lead to unreproducible methods and analyses. While important to start working in data analysis as fast as possible in the early beginning of a pandemic, it is also crucial that the models and analysis developed with the first data are validated a posteriori and confirmed with the updated data. It is thus fundamental that the subsequent datasets follow the same metadata and preferably are more complete and with less inconsistencies and errors.

Quality of healthcare data can be improved through several strategies. First, data entry processes must be simplified, avoiding duplications and reusing the data already in the system, since the need to input the same information in different systems is time-consuming, frustrating for the user, and can negatively impact both data completeness and accuracy. Data interoperability can also be a powerful approach to minimise the number of interactions with the system.²⁴ Second, data need to be constantly monitored and tracked²⁵: organisations must develop processes to evaluate data patterns, and establish report systems based on data quality metrics. Even before data curation, simple validation procedures and rules in information systems can help detect and prevent many errors (ie, male patients classified as 'pregnant', or a patient aged 134 years old) and inconsistencies, and improve data completeness.

Finally, we need to establish the value proposition for both creators and observers.²⁶ This includes ensuring that healthcare providers understand the importance of data, receive feedback about their analysis and how it may improve both the assistance to the patient and the whole organisation, and have received adequate training for better performance.

The adoption of these strategies should pave the way to high-quality, accurate healthcare datasets that can generate accurate knowledge to timely inform health policies, and the readaptation of healthcare systems to new challenges.

We acknowledge that our study has some limitations. One of such limitations is the lack of clarification by the data provider on the issues found in the datasets. In fact, despite repeated requests, we did not receive from DGS complete answers that could clarify the issues described in the manuscript. Therefore, the analysis of the Portuguese surveillance data quality was done exclusively with the analysis of the databases provided by DGS to researchers and with our external knowledge about how the information flows from the moment the data are introduced by health professionals until the dataset can be used for data analysis. Another limitation is the fact that we only studied the quality issues of COVID-19 data from one country, Portugal. However, our results seem to be in line with the findings of Ashofteh and Bravo⁷ who analysed and compared the quality of official datasets available for COVID-19, including data from the Chinese Center for Disease Control and Prevention, the WHO, and the European Centre for Disease Prevention and Control. In fact, they also found noticeable and increasing measurement errors in the three datasets as the pandemic outbreak expanded and more countries contributed data for the official repositories.

CONCLUSION

We describe some important quality issues of the Portuguese COVID-19 surveillance datasets, relevant enough to force the discussion about the validity of the published findings arising from these and similar data.

The availability of official data by the National Health Authorities to researchers is an enormous asset, allowing data analysis, modelling and prediction that may support better decisions for the patient and the community as a whole. However, to fully embrace this potential, it is crucial that these data are accurate and reliable.

System interoperability would be needed to allow the connection with all the different EHRs that are in use in Portugal. Most EHRs collect data using unstructured data fields that would be difficult to correctly extract to a form like the one in the National Surveillance Systems.

It also urges to define and implement major improvements in the processes and systems of surveillance datasets: simplification of data entry processes, constant monitoring of data, raising awareness of healthcare providers for the importance of good data and providing them adequate training.

Data curation processes, capitalising on effective and multidisciplinary collaborations between healthcare providers and data analysts, play a critical role to ensure minimum quality standards. Once these processes are fully optimised, the reliability of results and the quality of the scientific evidence produced can be greatly improved.

Author affiliations

¹Department of Community Medicine, Information and Health Decision Sciences (MEDCIDS), Faculty of Medicine, University of Porto, Porto, Portugal

²Centre for Health Technology and Services Research (CINTESIS), Faculty of Medicine, University of Porto, Porto, Portugal

³Patient Safety Translational Research Centre, Institute of Global Health Innovation, Imperial College London, London, UK

⁴Escola Superior de saúde da Cruz Vermelha Portuguesa, Lisbon, Portugal

⁵Porto Pharmacovigilance Centre, Faculty of Medicine, University of Porto, 4200-450 Porto, Portugal

Contributors CC-S conceptualised the study, contributed to design, analysis and interpretation of data, and drafted the manuscript. CC-S was also the guarantor and accepts full responsibility for the finished work and the conduct of the study, had access to the data, and controlled the decision to publish. ALN, RC, PS, MM-S, AF, IR-V, PPR, AC-P, AMP and JAF made substantial contributions to the conception and design of the study and revised the draft critically for important intellectual content. TSH made substantial contributions to the analysis and interpretation of data and revised the draft critically for important intellectual content.

Funding This work was supported by National Funds through FCT - Fundação para a Ciência e a Tecnologia, I.P., within CINTESIS, R&D Unit (reference UIDB/4255/2020).

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval The data used in this work were anonymised and made available by the Portuguese Directorate-General of Health (DGS), under the scope of article 39th of the decree law 2-B/2020, from 2 April. The study was approved on 17 April 2020 by the Health Ethics Committee of Centro Hospitalar Universitário de São João and Faculty of Medicine, University of Porto (number not available).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. The data related to this study (deidentified COVID-19 cases information) were made available by Portuguese Directorate-General of Health to authors upon request and after submission of a research proposal and documented approval by an ethical committee. All data are available from the corresponding author on a reasonable request and after authorisation from Portuguese Directorate-General of Health.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Cristina Costa-Santos <http://orcid.org/0000-0002-7109-1101>

Ana Luisa Neves <http://orcid.org/0000-0002-7107-7211>

Paulo Santos <http://orcid.org/0000-0002-2362-5527>

Matilde Monteiro-Soares <http://orcid.org/0000-0002-4586-2910>

REFERENCES

- Morgan O. How decision makers can use quantitative approaches to guide outbreak responses. *Philos Trans R Soc Lond B Biol Sci* 2019;374:20180365.
- Xu B, Kraemer MUG, Gutierrez B, Open COVID-19 Data Curation Group. Open access epidemiological data from the COVID-19 outbreak. *Lancet Infect Dis* 2020;20:534.

- 3 Yozwiak NL, Schaffner SF, Sabeti PC. Data sharing: make outbreak research open access. *Nature* 2015;518:477–9.
- 4 German RR, Lee LM, Horan JM, *et al.* Updated guidelines for evaluating public health surveillance systems: recommendations from the guidelines Working group. *MMWR Recomm Rep* 2001;50:1–35; quiz CE1–7.
- 5 Alonso V, Santos JV, Pinto M, *et al.* Health records as the basis of clinical coding: is the quality adequate? A qualitative study of medical coders' perceptions. *Health Inf Manag* 2020;49:28–37.
- 6 Chen H, Hailey D, Wang N, *et al.* A review of data quality assessment methods for public health information systems. *Int J Environ Res Public Health* 2014;11:5170–207.
- 7 Ashofteh A, Bravo JM. A study on the quality of novel coronavirus (COVID-19) official datasets. *Stat J IAOS* 2020;36:291–301.
- 8 Boes L, Houareau C, Altmann D, *et al.* Evaluation of the German surveillance system for hepatitis B regarding timeliness, data quality, and simplicity, from 2005 to 2014. *Public Health* 2020;180:141–8.
- 9 Tosti ME, Longhi S, de Waure C, *et al.* Assessment of timeliness, representativeness and quality of data reported to Italy's national integrated surveillance system for acute viral hepatitis (SEIEVA). *Public Health* 2015;129:561–8.
- 10 Yazidi R, Aissi W, Bouguerra H, *et al.* Evaluation of the influenza-like illness surveillance system in Tunisia, 2012–2015. *BMC Public Health* 2019;19:694.
- 11 Visa TI, Ajumobi O, Bamgboye E, *et al.* Evaluation of malaria surveillance system in Kano state, Nigeria, 2013–2016. *Infect Dis Poverty* 2020;9:15.
- 12 Wolkewitz M, Puljak L. Methodological challenges of analysing COVID-19 data during the pandemic. *BMC Med Res Methodol* 2020;20:81.
- 13 Direção Geral da Saúde. Comunicado: Casos de infeção POR novo Coronavírus (COVID-19), 2020. Available: <https://covid19.min-saude.pt/wp-content/uploads/2020/03/Atualiza%C3%A7%C3%A3o-de-02032020-1728.pdf> [Accessed 17 Aug 2020].
- 14 Carta aberta AO Conselho Nacional de Saúde Pública: Um contributo pessoal acerca da epidemia de Covid-19, em Portugal, 2020. Available: https://sigarra.up.pt/fmup/pt/noticias_geral.noticias_cont?p_id=F307210300/CartaAberta_COVID19_11.03.2020_.pdf [Accessed 17 Aug 2020].
- 15 Direção Geral da Saúde. COVID-19: Disponibilização de Dados, 2020. Available: <https://covid19.min-saude.pt/disponibilizacao-de-dados/> [Accessed 11 Aug 2020].
- 16 Direção Geral da Saúde. COVID metadata, 2020. Available: https://covid19.min-saude.pt/wp-content/uploads/2020/04/PT_COVID19_metadata-1.pdf [Accessed 11 Aug 2020].
- 17 Direção Geral da Saúde. Relatório de Situação - Informação publicada diariamente, 2020. Available: <https://covid19.min-saude.pt/relatorio-de-situacao/> [Accessed 11 Aug 2020].
- 18 Costa-Santos C, Ribeiro-Vaz I, Monteiro-Soares M. The hidden factor-low quality of data is a major peril in the identification of risk factors for COVID-19 deaths: a comment on Nogueira, P.J., *et al.* "The role of health preconditions on COVID-19 deaths in Portugal: evidence from surveillance data of the first 20293 infection cases". *J. Clin. Med.* 2020, 9, 2368. *J Clin Med* 2020;9:3442.
- 19 Nogueira PJ, de Araújo Nobre M, Costa A, *et al.* The role of health preconditions on COVID-19 deaths in Portugal: evidence from surveillance data of the first 20293 infection cases. *J Clin Med* 2020;9:2368.
- 20 Sa-Sousa A, Morais-Almeida M, Azevedo LF, *et al.* Prevalence of asthma in Portugal - The Portuguese National Asthma Survey. *Clin Transl Allergy* 2012;2:15.
- 21 Quinaz Romana G, Kislaya I, Salvador MR, *et al.* [Multimorbidity in Portugal: Results from The First National Health Examination Survey]. *Acta Med Port* 2019;32:30–7.
- 22 Peixoto R, Viera V, Aguar A. COVID-19: determinants of hospitalization, ICU and death among 20,293 reported cases in Portugal. *medRxiv* 2020.
- 23 Froes MT, Neves BD, Martins B. Comparison of multimorbidity in COVID-19 infected and general population in Portugal. *medRxiv* 2020.
- 24 D'Amore J, Bouhaddou O, Mitchell S, *et al.* Interoperability progress and remaining data quality barriers of certified health information technologies. *AMIA Annu Symp Proc* 2018;2018:358–67.
- 25 Chen H, Hailey D, Wang N, *et al.* A review of data quality assessment methods for public health information systems. *Int J Environ Res Public Health* 2014;11:5170–207.
- 26 IOM Roundtable on Value & Science-Driven Care, Institute of Medicine. Integrating Research and Practice: Health System Leaders Working Toward High-Value Care: Workshop Summary. In: *Continuously learning health care: the value proposition*. Washington (DC): National Academies Press (US), 2015. <https://www.ncbi.nlm.nih.gov/books/NBK284656/>